

Introduction:

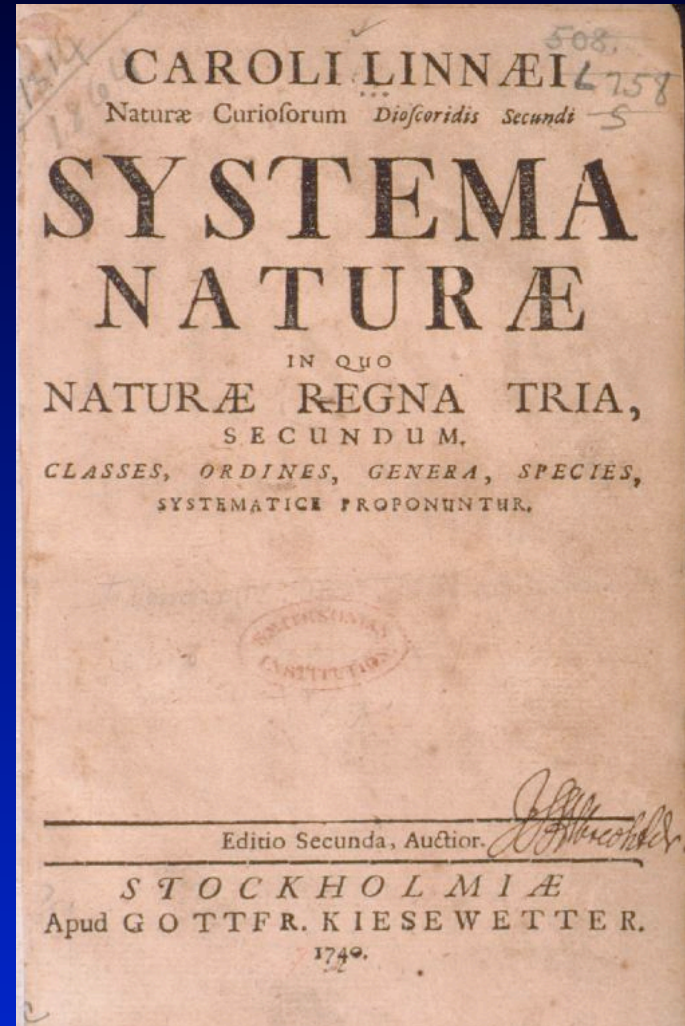
Theory of Evolution Bioinformatics

Anders Gorm Pedersen
Molecular Evolution Group

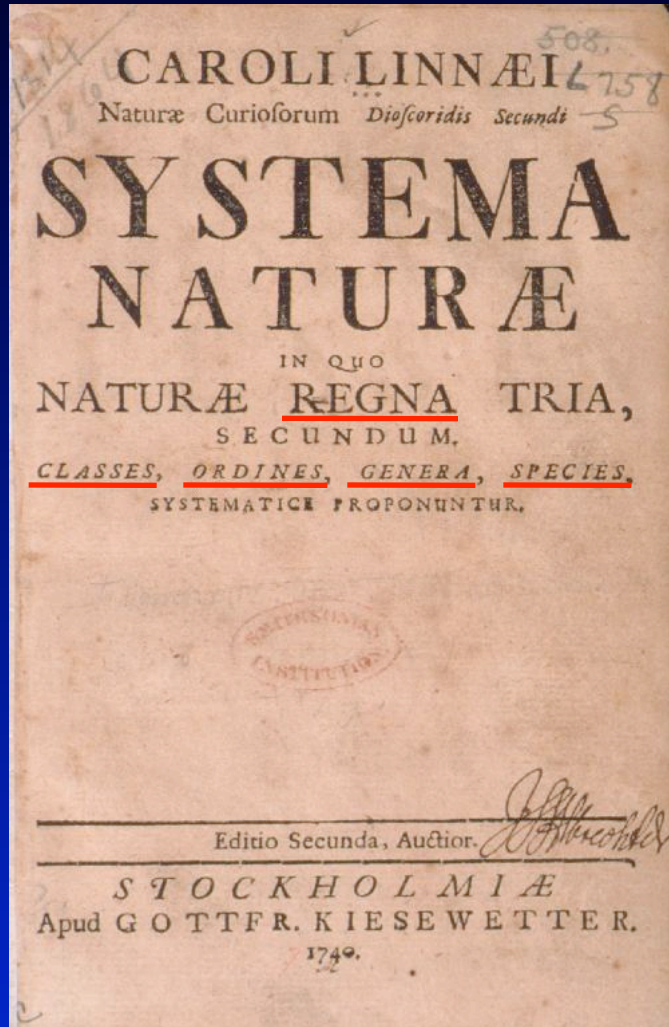
Classification: Linnaeus



Carl Linnaeus
1707-1778

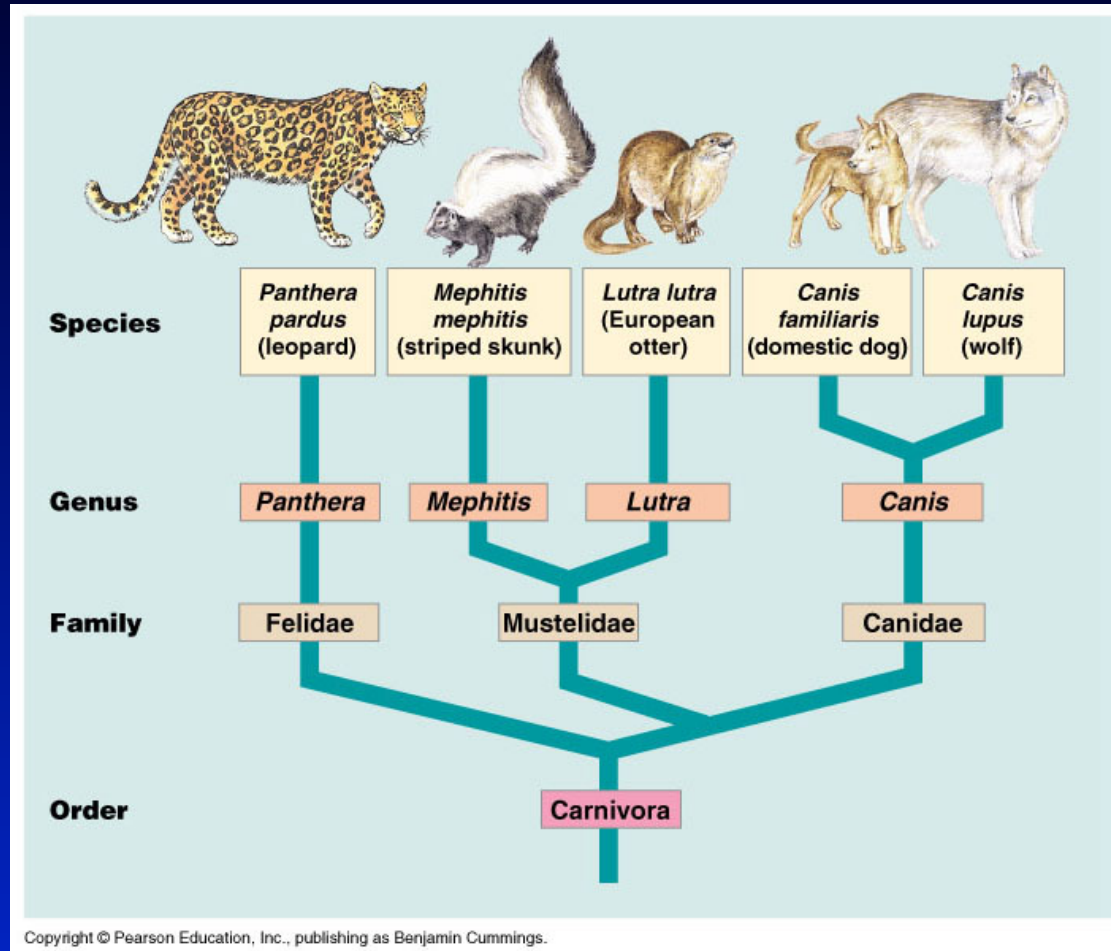


Classification: Linnaeus

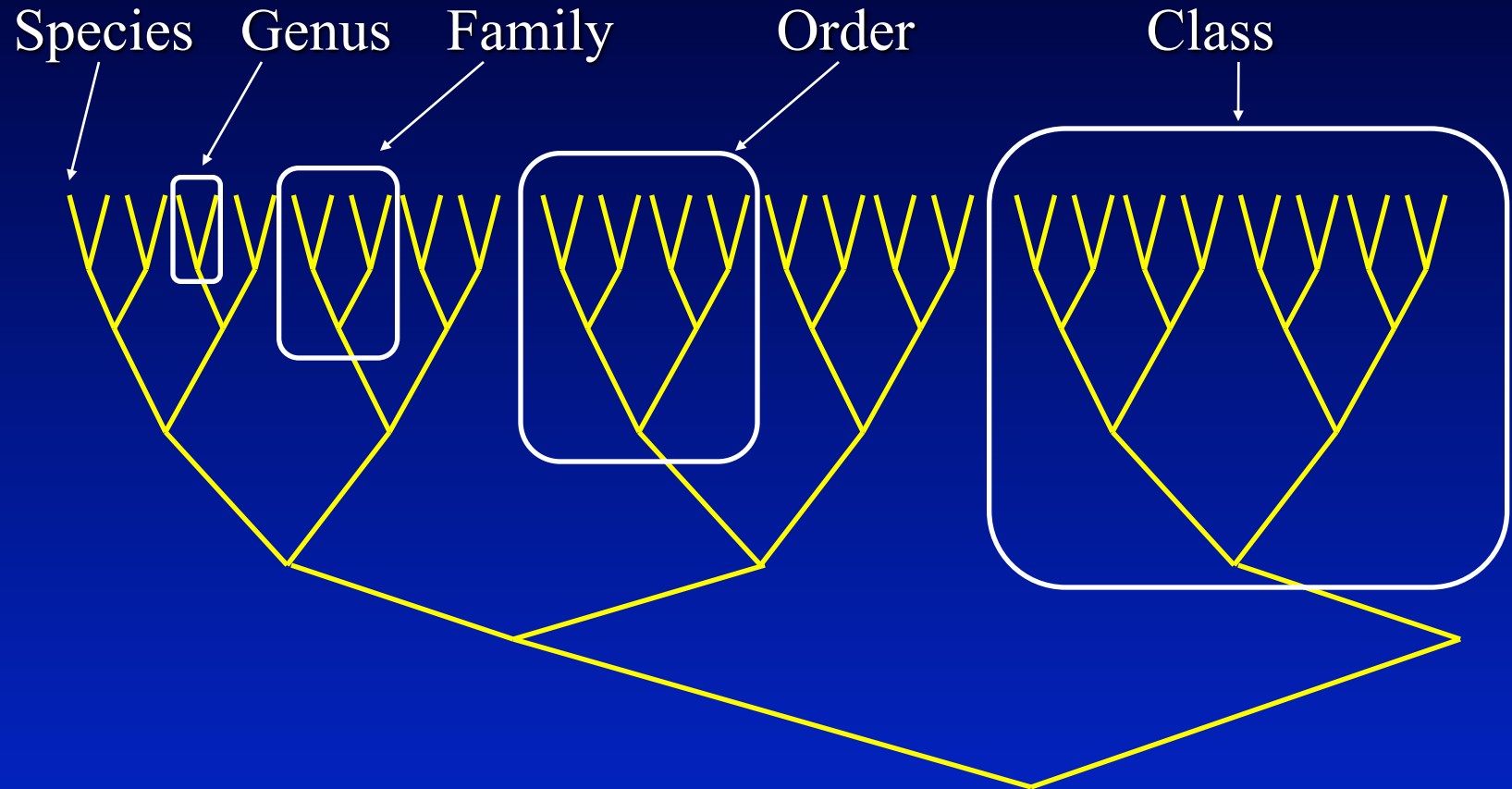


- Hierarchical system
 - Kingdom
 - Phylum
 - Class
 - Order
 - Family
 - Genus
 - Species

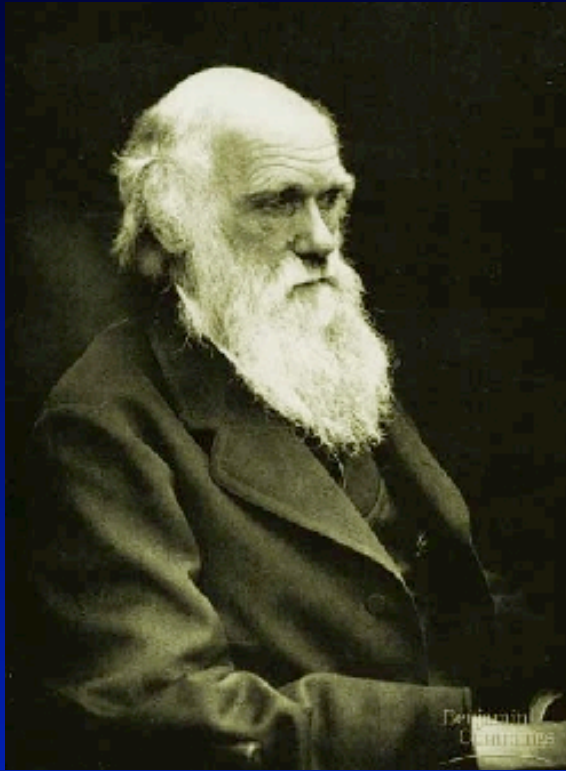
Classification depicted as a tree



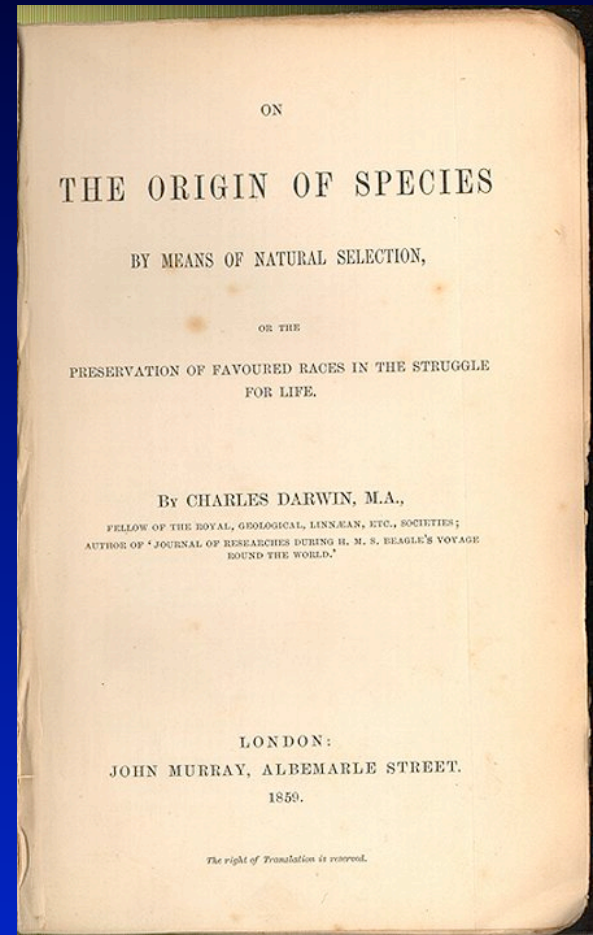
Classification depicted as a tree



Theory of evolution

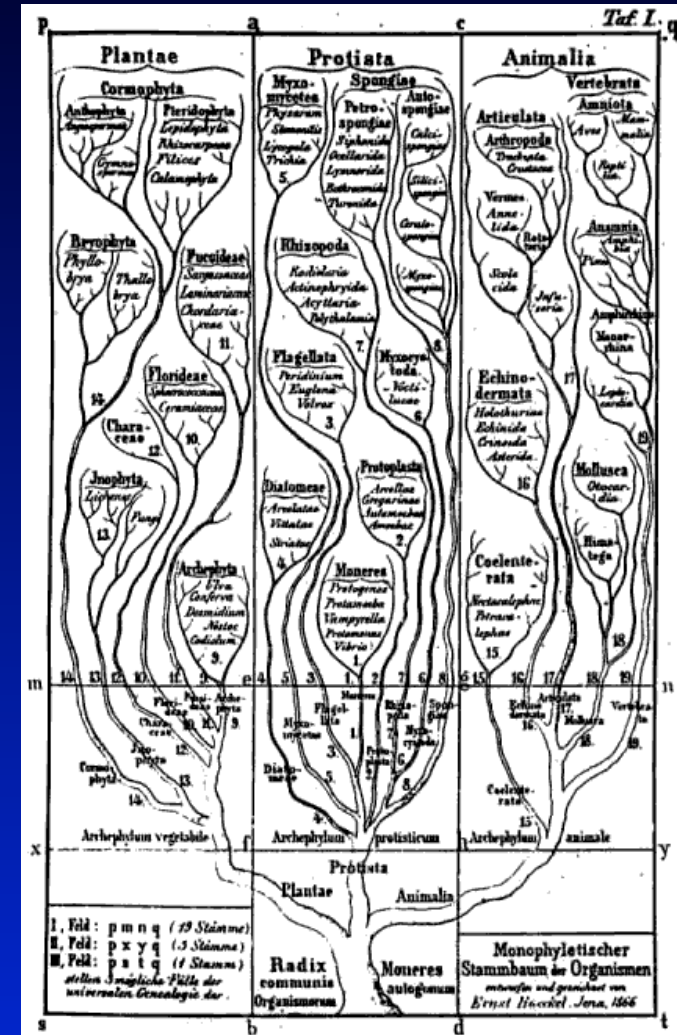


Charles Darwin
1809-1882



Phylogenetic basis of systematics

- **Linnaeus:**
Ordering principle is God.
- **Darwin:**
Ordering principle is shared descent from common ancestors.
- Today, systematics is explicitly based on phylogeny.

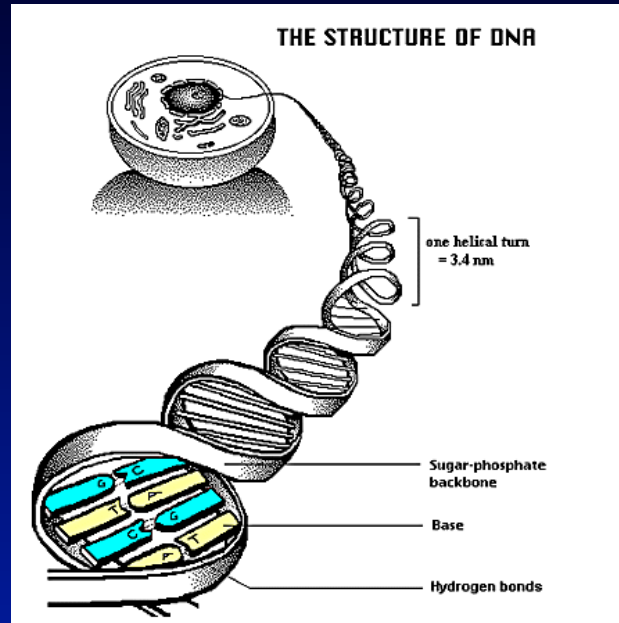


Natural Selection: Darwin's four postulates

- More young are produced each generation than can survive to reproduce.
- Individuals in a population vary in their characteristics.
- Some differences among individuals are based on genetic differences.
- Individuals with favorable characteristics have higher rates of survival and reproduction.
- Evolution by means of natural selection
- Presence of "design-like" features in organisms:
- Quite often features are there "for a reason"

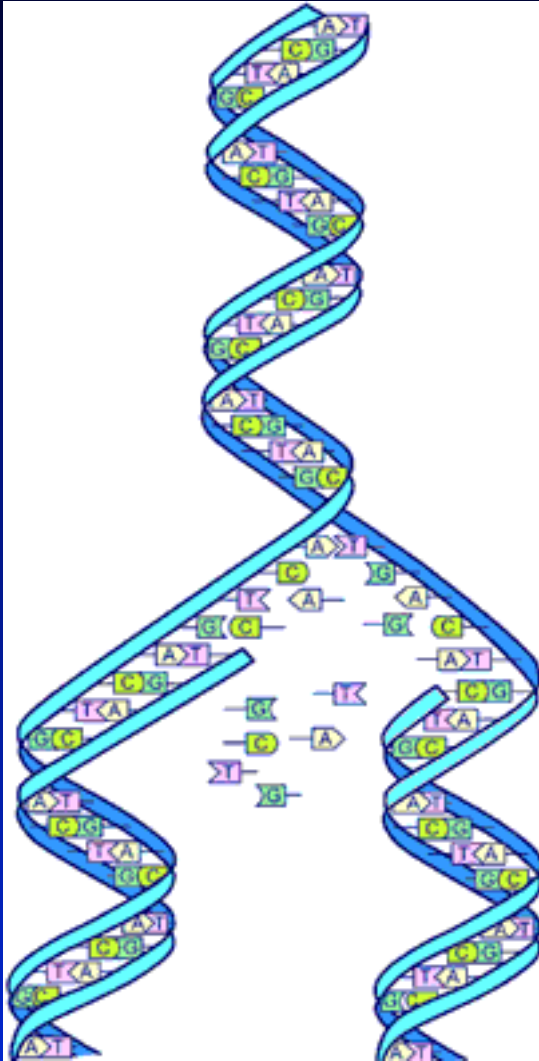
Bioinformatics

What is bioinformatics?



Bioinformatics: computational analysis of biological data

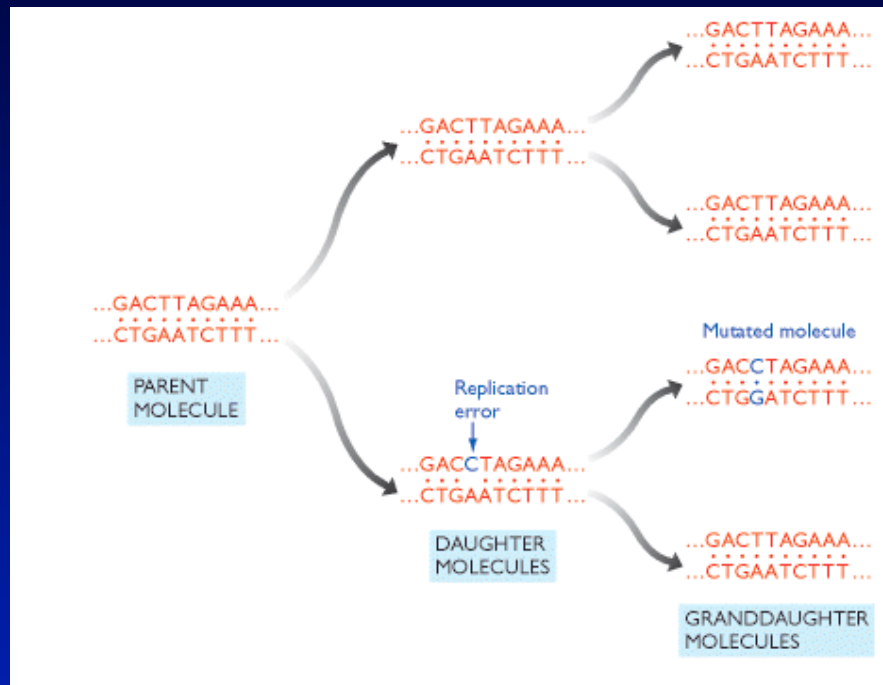
Molecular Basis for Heredity: DNA



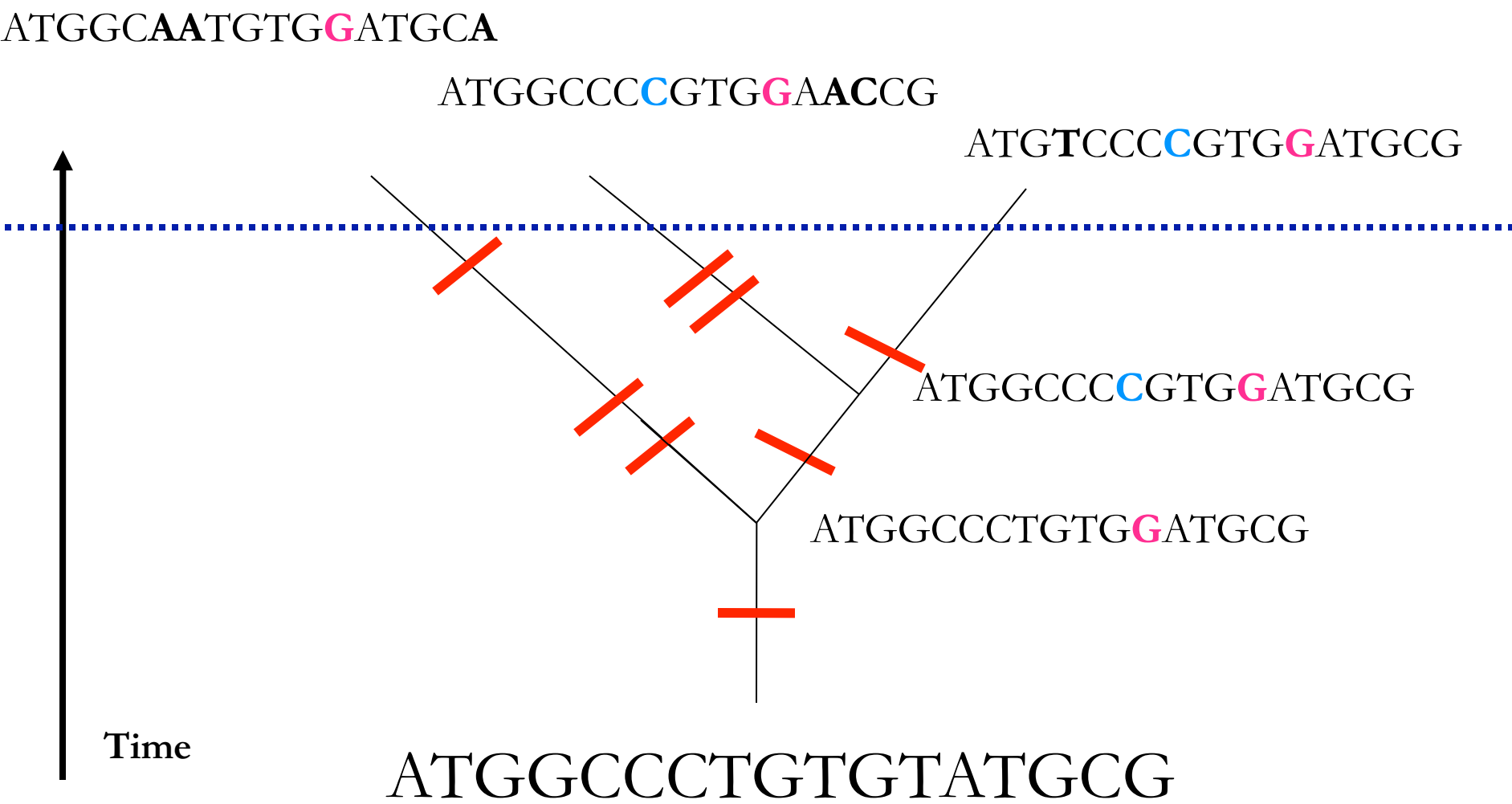
A always paired with T

C always paired with G

Molecular Basis for Evolution: DNA Mutation

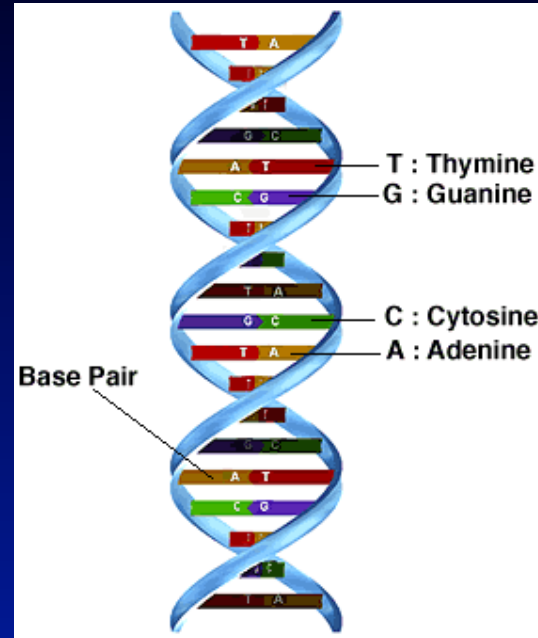
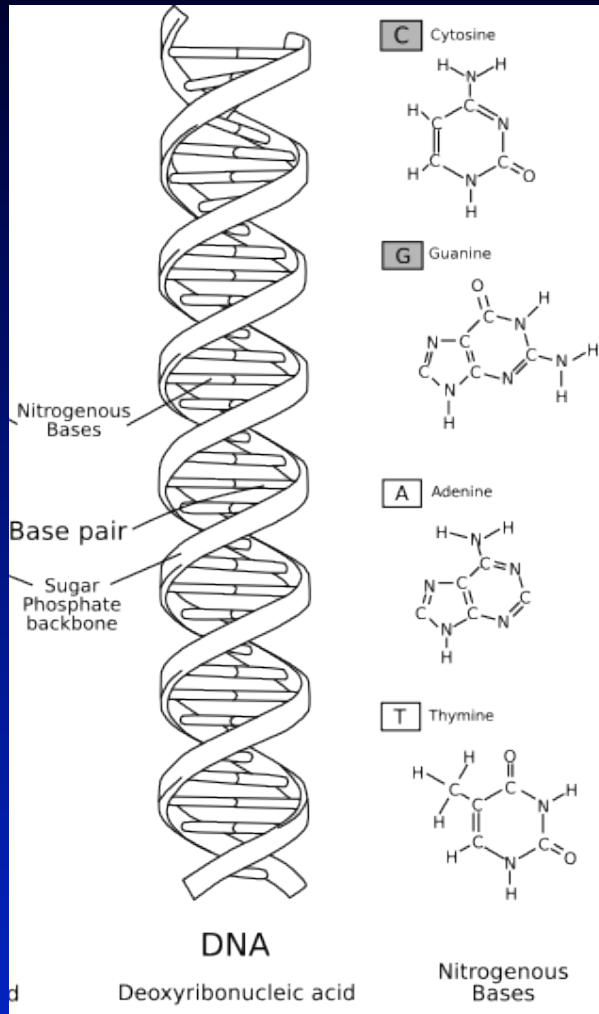


A history of mutations



- Species1: ATGGC**AA**TGTG**G**ATGCA**A**
 - Species2: ATGGCCCC**C**GTG**G**A**AC**CG
 - Species3: ATG**T**CCCC**C**GTG**G**ATGCG
- Diagram illustrating DNA alignment for three species. The sequences are aligned, and the number of matches (indicated by brackets) is shown for each species relative to a reference sequence (Species1).
- Species1: 6 matches (indicated by a bracket labeled 6)
- Species2: 3 matches (indicated by a bracket labeled 3)
- Species3: 3 matches (indicated by a bracket labeled 3)
- The total number of matches across all species is 5 (indicated by a bracket labeled 5).

Symbolic representation of DNA structure

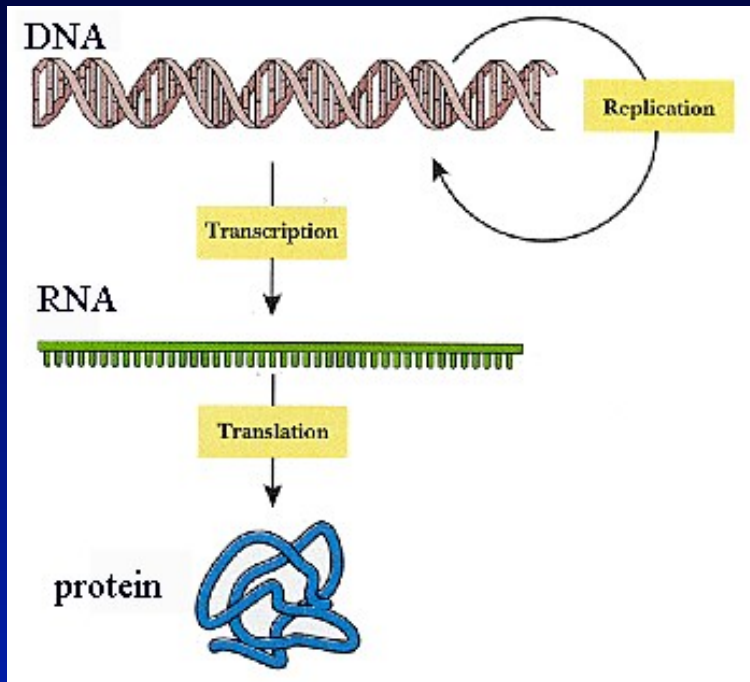


- DNA molecule is a linear polymer
- Structure can be represented as string of 4 symbols: ACTG
- These “sequences” can be analyzed mathematically/linguistically

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

[illegible]

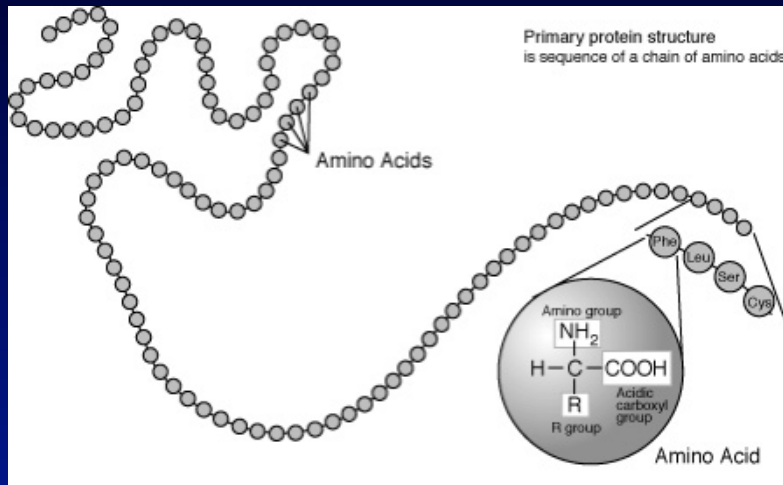
DNA --> RNA --> protein



Standard Genetic Code

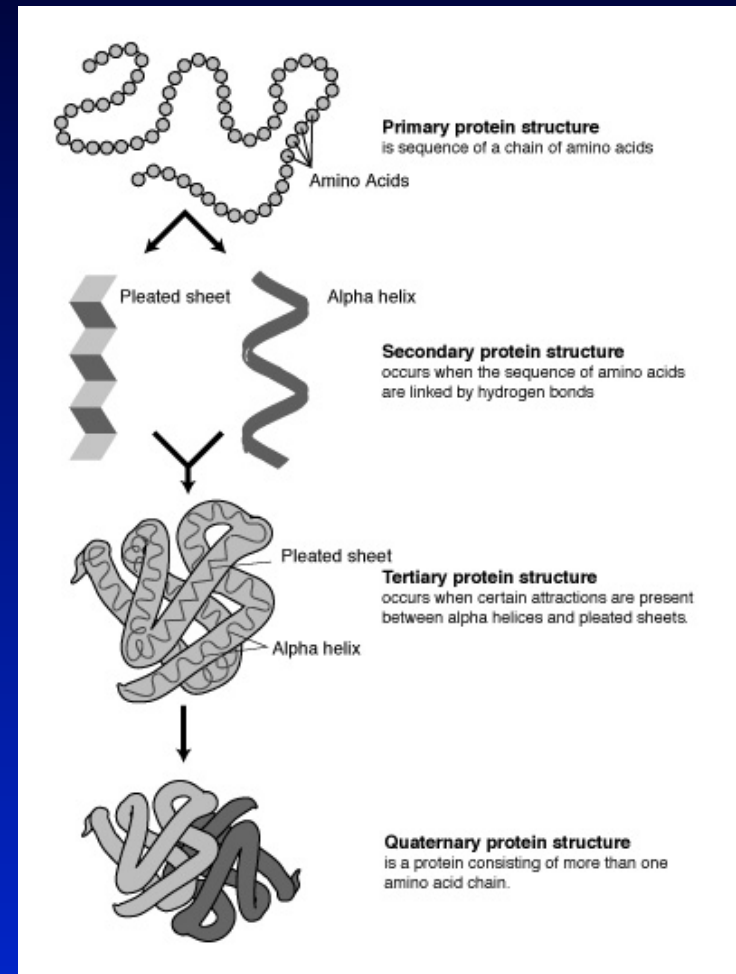
	T			C			A			G			
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C	T
	TTC	Phe	F	TCC	Ser	S	TAC	Tyr	Y	TGC	Cys	C	C
	TTA	Leu	L	TCA	Ser	S	TAA	Och *		TGA	Opa *		A
	TTG	Leu	L	TCG	Ser	S	TAG	Amb *		TGG	Trp	W	G
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R	T
	CTC	Leu	L	CCC	Pro	P	CAC	His	H	CGC	Arg	R	C
	CTA	Leu	L	CCA	Pro	P	CAA	Gln	Q	CGA	Arg	R	A
	CTG	Leu	L	CCG	Pro	P	CAG	Gln	Q	CGG	Arg	R	G
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S	T
	ATC	Ile	I	ACC	Thr	T	AAC	Asn	N	AGC	Ser	S	C
	ATA	Ile	I	ACA	Thr	T	AAA	Lys	K	AGA	Arg	R	A
	ATG	Met	M	ACG	Thr	T	AAG	Lys	K	AGG	Arg	R	G
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G	T
	GTC	Val	V	GCC	Ala	A	GAC	Asp	D	GGC	Gly	G	C
	GTA	Val	V	GCA	Ala	A	GAA	Glu	E	GGA	Gly	G	A
	GTG	Val	V	GCG	Ala	A	GAG	Glu	E	GGG	Gly	G	G

Symbolic representation of protein structure



- Proteins are linear polymers
- Built from 20 amino acids
- Can be represented as string of 20 symbols

ACDEFGHIKLMNPQRSTVWY



NCBI databases

The screenshot displays the NCBI (National Center for Biotechnology Information) website. At the top, the browser address bar shows the URL <http://www.ncbi.nlm.nih.gov/>. Below the address bar is a navigation menu with links to various resources like Flu, CBS, DTU, Literature, NCBI, Science, Wikipedia, Mac, Translation, Plagiarism, Wifi, TV, Fun, Blogs, and Exercise. The main header area includes the NCBI logo, a search bar with the text "human globin", and a "Search" button. The page is divided into several sections: "Resources" on the left with a list of links; "Welcome to NCBI" in the center with a description of the center's mission; "Genome Reference Consortium" with a brief description and a list of links; "How To..." with a list of instructions; "NLN/NCBI H1N1 Flu Resources" at the bottom; "Popular Resources" on the right with a list of links; and "NCBI News" with a list of recent news items.

Resources

- NCBI Home
- All Resources (A-Z)
- Literature
- DNA & RNA
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structures
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[More about the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS](#)

Genome Reference Consortium

Formed to improve human and mouse reference assemblies, GRC will fix loci misrepresented in reference assembly, fill remaining gaps, and make alternate representations of complex loci.

How To...

- Obtain the full text of an article
- Retrieve all sequences for an organism or taxon
- Find a homolog for a gene in another organism
- Find genes associated with a phenotype or disease
- Design PCR primers and check them for specificity
- Find the function of a gene or gene product
- Determine conserved synteny between the genomes of two organisms

[See all ...](#)

NLM/NCBI H1N1 Flu Resources

Popular Resources

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

NCBI News

November and October News 02 Dec 2009
Featured: New Discovery-oriented PubMed and NCBI Homepage, T...

NCBI News - September 2009 05 Oct 2009
The September 2009 issue of the NCBI News is available ...

NCBI News - August 2009 19 Aug 2009
The August 2009 issue of the NCBI News is available online. ...

[More...](#)

NCBI databases: Genbank feature table

Nucleotide - Homo sapiens hemoglobin, gamma A (HBG1), mRNA

Entrez Gene record to access additional publications.
COMPLETENESS: full length.

```
FEATURES             Location/Qualifiers
     source             1..444
                        /organism="Homo sapiens"
                        /mol_type="mRNA"
                        /db_xref="taxon:9606"
                        /chromosome="11"
                        /map="11p15.5"
     gene               <1..>444
                        /gene="HBG1"
                        /gene_synonym="HBGA; HBGR; HSGGL1; PRO2979"
                        /note="hemoglobin, gamma A"
                        /db_xref="GeneID:3047"
                        /db_xref="HGNC:4831"
                        /db_xref="HPRD:00789"
                        /db_xref="MIM:142200"
     exon               <1..92
                        /gene="HBG1"
                        /gene_synonym="HBGA; HBGR; HSGGL1; PRO2979"
                        /inference="alignment:Splign"
                        /number=1
     CDS                1..444
                        /gene="HBG1"
                        /gene_synonym="HBGA; HBGR; HSGGL1; PRO2979"
                        /note="hemoglobin gamma-a chain; hemoglobin, gamma,
                        regulator of; gamma globin; gamma A hemoglobin"
                        /codon_start=1
                        /product="A-gamma globin"
                        /protein_id="NP_000550.2"
                        /db_xref="GI:28302131"
                        /db_xref="CCDS:CCDS7754.1"
                        /db_xref="GeneID:3047"
                        /db_xref="HGNC:4831"
                        /db_xref="HPRD:00789"
                        /db_xref="MIM:142200"
                        /translation="MCHFTEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFD
                        SFGNLSASAIMGNPKVKAHGKKVLTSLGDATKHLDDLKGTFAQLSELHCDKLHVDPE
                        NFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTAVASALSSRYH"
     exon               93..315
                        /gene="HBG1"
                        /gene_synonym="HBGA; HBGR; HSGGL1; PRO2979"
                        /inference="alignment:Splign"
                        /number=2
     STS                194..369
                        /gene="HBG1"
```


NCBI databases: fasta format

NCBI Nucleotide - Homo sapiens hemoglobin, gamma A (HBG1), mRNA

http://www.ncbi.nlm.nih.gov/nuccore/28302130?report=fasta&log\$=seqview&from=54&to=497

Search Nucleotide for [] Go Clear

Format: GenBank FASTA Graphics More Formats

Showing 444 bp region from base 54 to 497.

NCBI Reference Sequence: NM_000559.2

Homo sapiens hemoglobin, gamma A (HBG1), mRNA

>gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA
ATGGGTCATTTACAGAGGAGGACAAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGAGGAGAAACCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGG
CAACCTGTCTCTGCTCTGCCATCATGGGCAACCCAAAGTCAAGGCACATGGCAAGAGGTGCTGACT
TCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAAGTGCCT
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTGGCAAT
CCATTTGGGCAAGAATTCAACCCCTGAGGTGACGGCTTCTGGCAGAAGATGGTGACTGCAGTGGCCAGT
GCCCTGTCTCCAGATACCACTGA

Change Region Shown

☐ Whole sequence
☒ Selected Region
from: 54 to: 497
Update View

Customize View

Analyze This Sequence

- Run BLAST
- Pick Primers

Articles about the HBG1 gene

- Molecular analysis of gamma-globin promoters, HS-111 and [Hemoglobin. 2009]
- A genome-wide association identified the common genetic variant [Hum Genet. 2009]
- Expression of miR-210 during erythroid differentiation and induction [BMB Rep. 2009]

» See all...

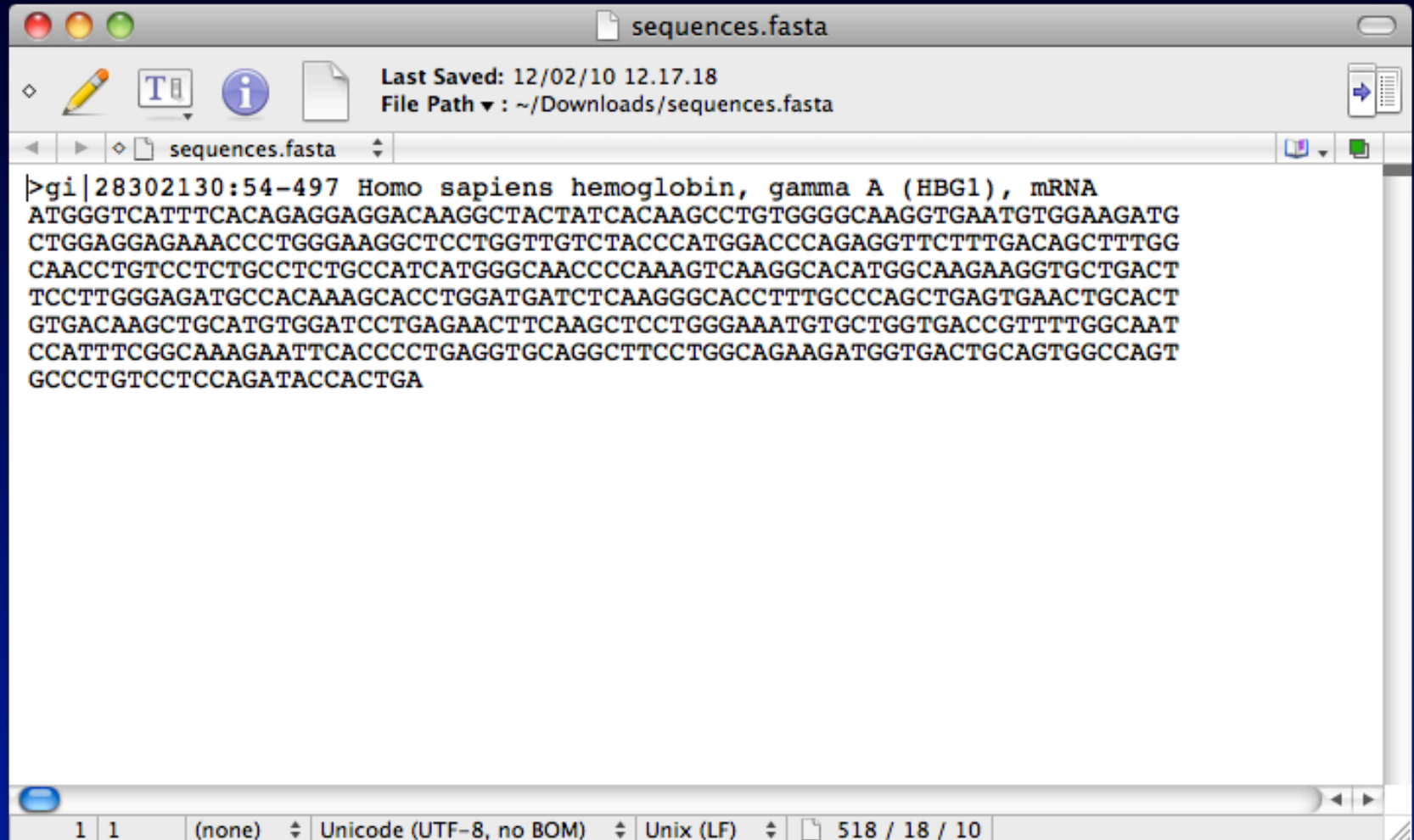
RefSeq Protein Product

See the reference protein sequence for A-gamma globin (NP_000550.2).

More about the HBG1 gene

The gamma-globin genes (HBG1 and HBG2)

FASTA file



sequences.fasta

Last Saved: 12/02/10 12.17.18
File Path ▾ : ~/Downloads/sequences.fasta

sequences.fasta

```
>gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA
ATGGGTCATTTACAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGG
CAACCTGTCTCTGCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACT
TCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAAGTGCAGT
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTTGGCAAT
CCATTTTCGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCTTGGCAGAAGATGGTGACTGCAGTGGCCAGT
GCCCTGTCCTCCAGATACCACTGA
```

1 1 (none) Unicode (UTF-8, no BOM) Unix (LF) 518 / 18 / 10